Deep Alternative Clustering

CAIML Summer School 2023

Lecture: Claudia Plant
Tutorial: Lukas Miklautz

Outline

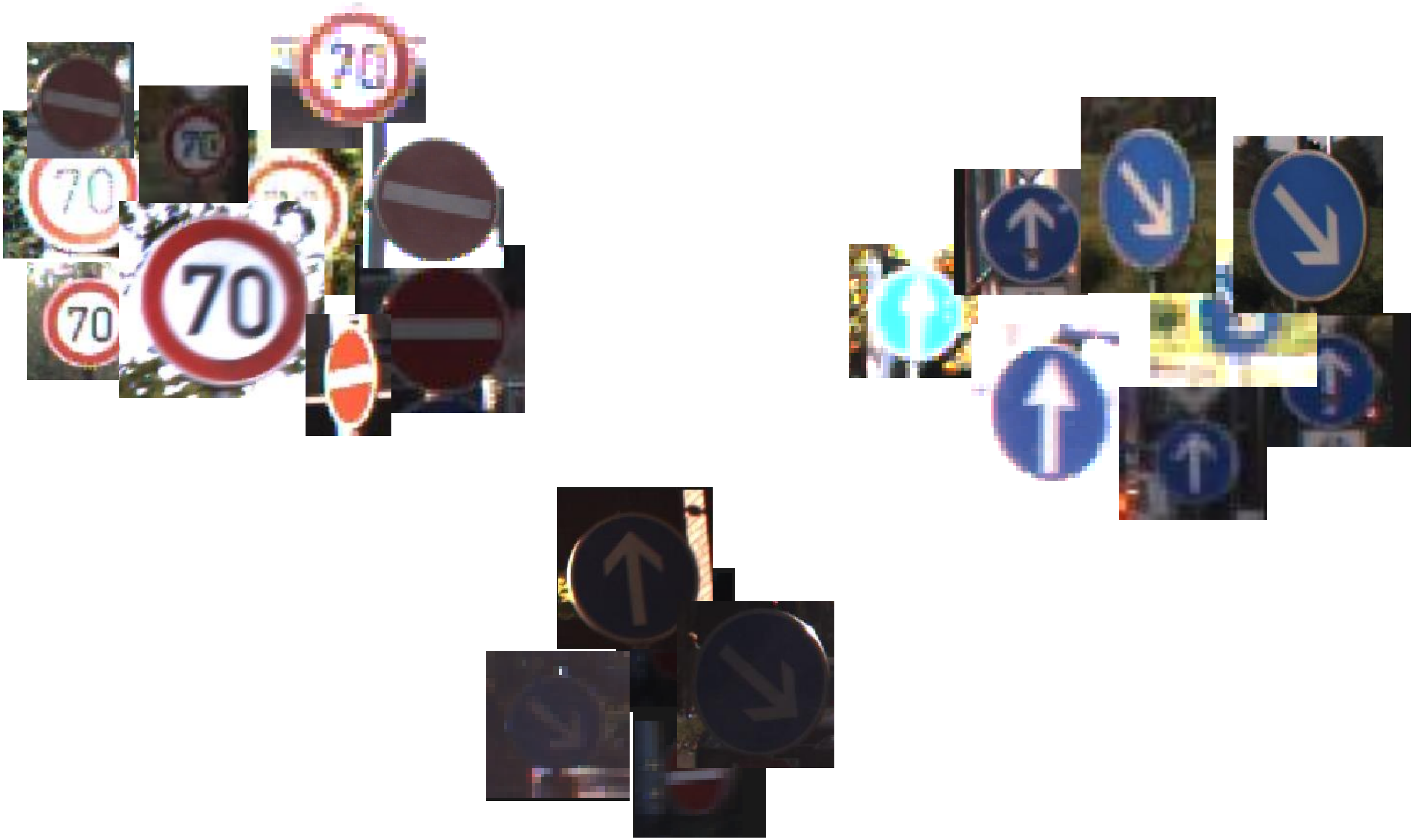
Clustering – find a meaningful grouping

universität wien
Fakultät für Informatik

Alternative Clustering

Goal:
Find all meaningful alternative clusterings.

Alternative Clustering

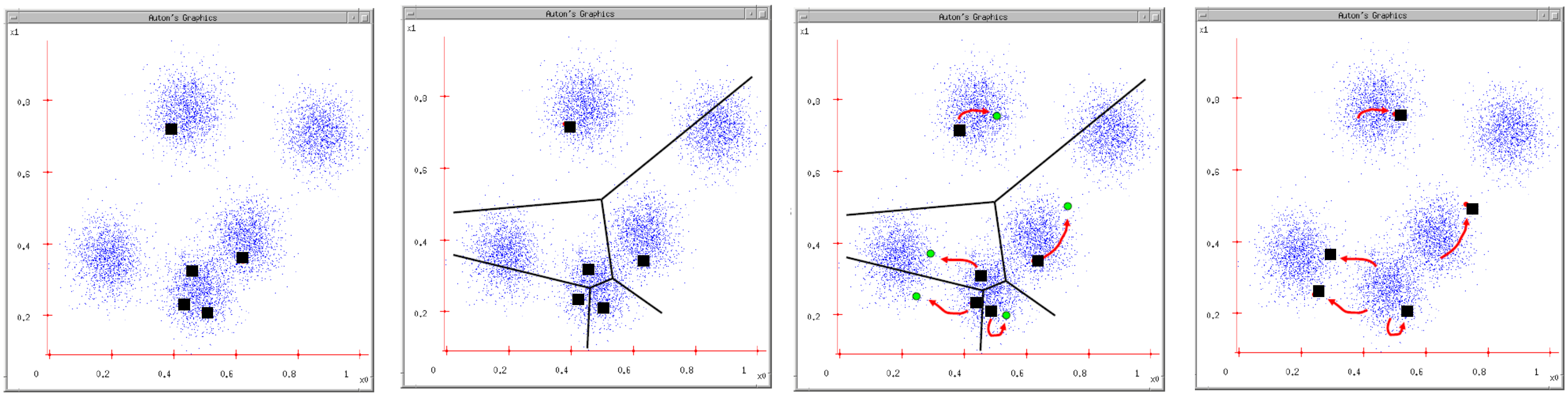


Outline

Fakultät für Informatik

Basis: K-Means

universität wien
Fakultät für Informatik



1) random **initialization** of the K cluster centers

2) **assignment** of the objects to the closest center

3) **update** of the centers

4) **iteration** of 2) and 3) until convergence

+ fast convergence,
+ well-defined objective function,
+ model.

$$\mathcal{F} = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} \left\| \boldsymbol{x} - \boldsymbol{\mu}_i \right\|^2$$

Multiple K-means Clusterings in Optimal Subspaces

$$\mathcal{F} = \sum_{j=1}^{S} \sum_{i=1}^{k_j} \sum_{\mathbf{x} \in C_{j,i}} \left\| P_j^\top V^\top \mathbf{x} - P_j^\top V^\top \boldsymbol{\mu}_{j,i} \right\|^2$$

Multiple K-means Clusterings in Optimal Subspaces

$$\mathcal{F} = \sum_{j=1}^{S} \sum_{i=1}^{k_j} \sum_{\mathbf{x} \in C_{j,i}} \left\| P_j^{\mathsf{T}} V^{\mathsf{T}} \mathbf{x} - P_j^{\mathsf{T}} V^{\mathsf{T}} \boldsymbol{\mu}_{j,i} \right\|^2$$

Multiple K-means Clusterings in Optimal Subspaces

$$\mathcal{F} = \boxed{\sum_{j=1}^{S}} \sum_{i=1}^{k_j} \sum_{\mathbf{x} \in C_{j,i}} \left\| P_j^{\mathsf{T}} V^{\mathsf{T}} \mathbf{x} - P_j^{\mathsf{T}} V^{\mathsf{T}} \boldsymbol{\mu}_{j,i} \right\|^2$$

- Now we consider S subspaces, each with k_j clusters

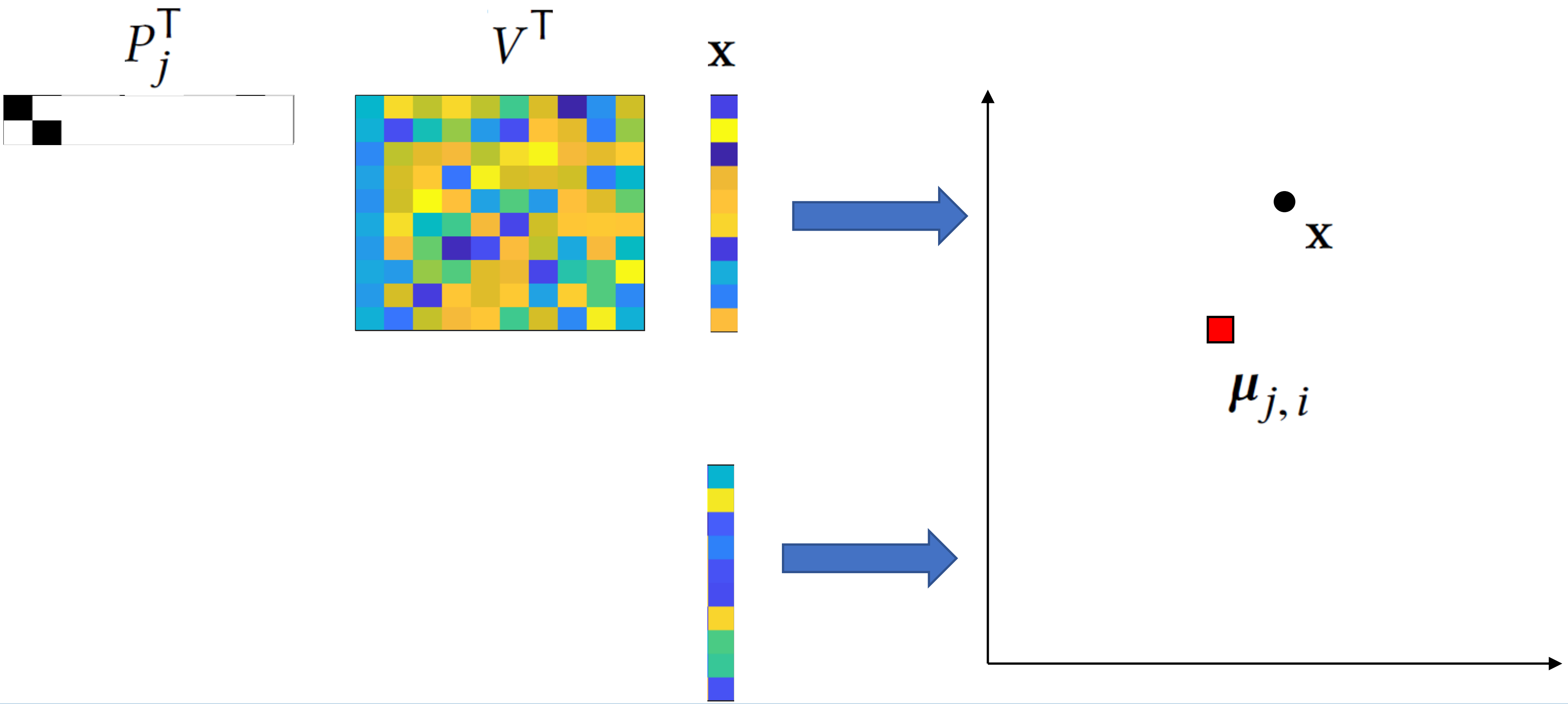
Multiple K-means Clusterings in Optimal Subspaces

$$\mathcal{F} = \boxed{\sum_{j=1}^{S}} \sum_{i=1}^{k_j} \sum_{\mathbf{x} \in C_{j,i}} \left\| P_j^\mathsf{T} V^\mathsf{T} \mathbf{x} - P_j^\mathsf{T} V^\mathsf{T} \boldsymbol{\mu}_{j,i} \right\|^2$$

- Now we consider *S* subspaces, each with k_j clusters
- V^T is a common an orthogonal transformation matrix
- P_j is a masking matrix that does the projection to Subspace *j*

Intuition: Rotation and Projection to Subspace

Assume the original data space is 10-dimensional and the subspace *j* is 2-dimensional.



Algorithm NR-K-Means: Initialization

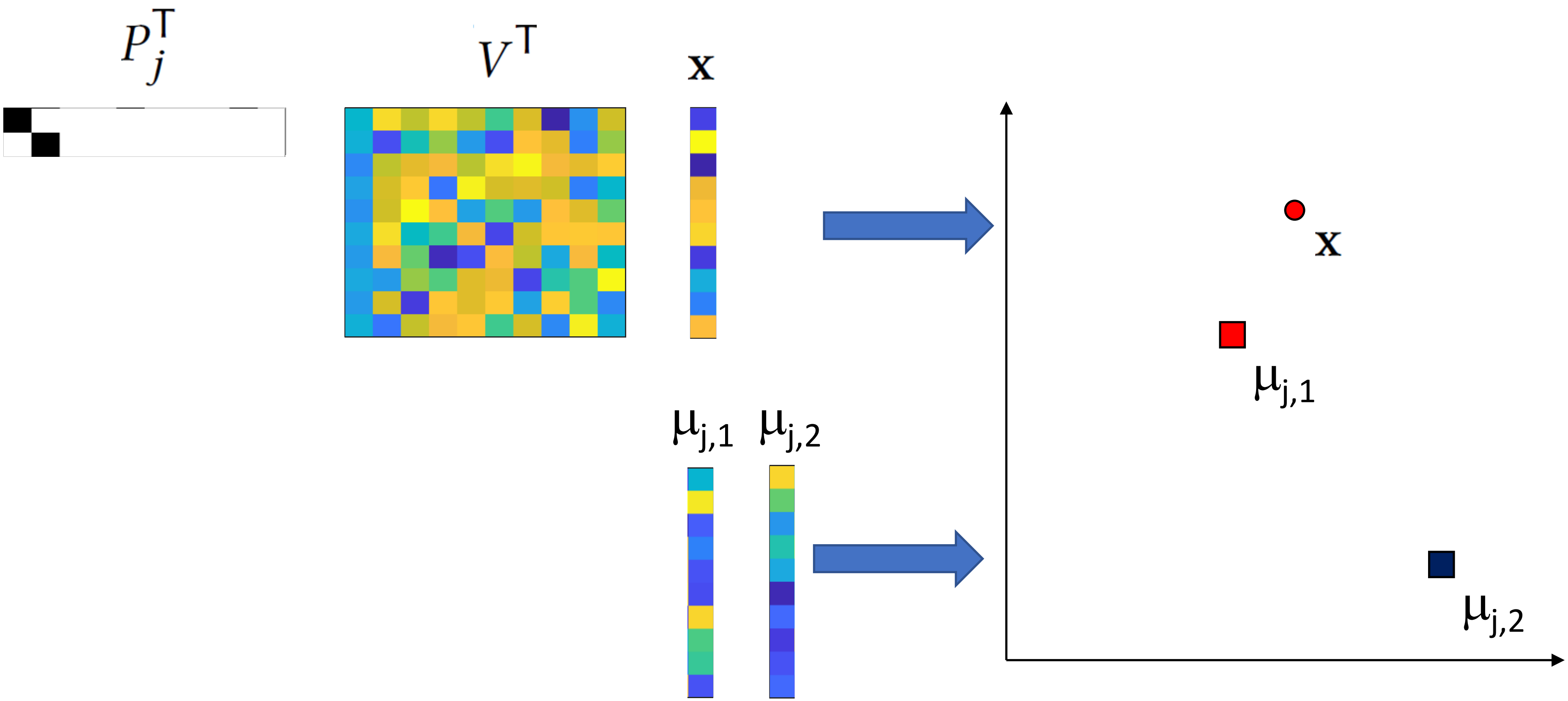
Input parameters:
number of subspaces S, number of clusters $k_1, \ldots, k_S$ in each subspace

- Intitialize V to a random orthogonal matrix
- Initialize the projection matrices P_j of size $d \times d/S$
- Initialize the cluster centers $m_{j,i}$ with a random data point

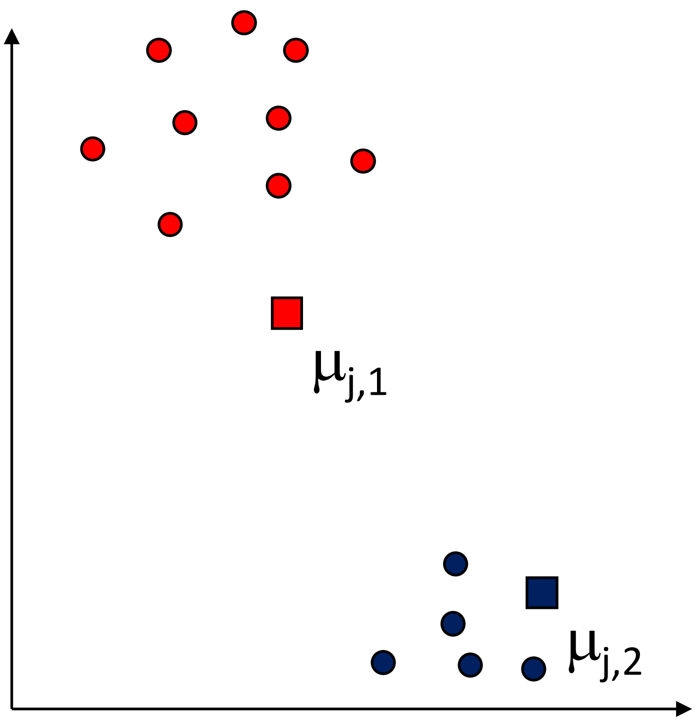
The algorithm will find automatically the optimal dimensionality for each subspace.

NR-K-Means: Assignment

For all subspaces: Project all points and all centers; assign each point to the closest center (Euclidean distance).



NR-K-Means: Update of the Cluster Centers

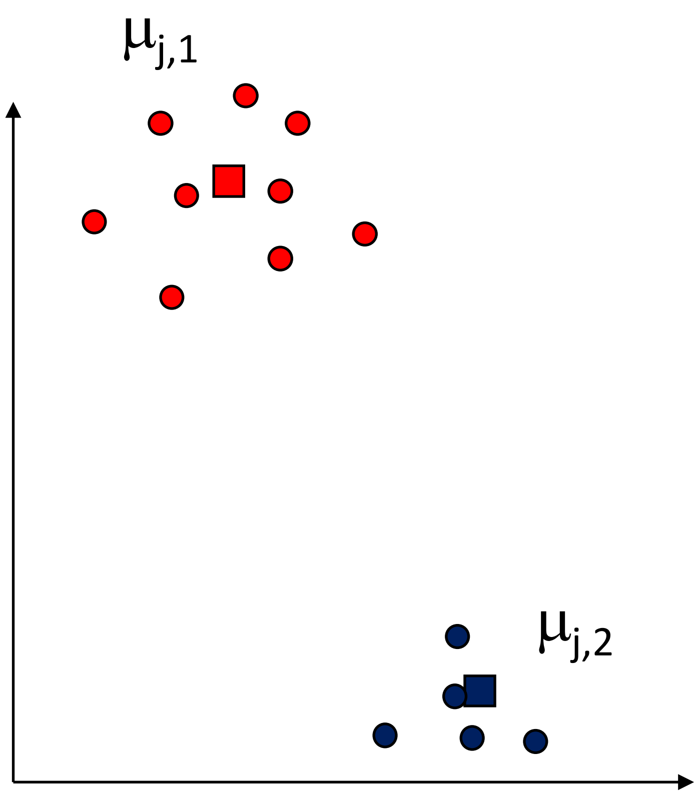


As in classical K-Means:
The cluster center is the mean of the associated points.

$$\boldsymbol{\mu}_{j,i} = \frac{1}{|C_{j,i}|} \sum_{\mathbf{x} \in C_{j,i}} \mathbf{x}$$

NR-K-Means: Update of the Cluster Centers

universität wien
Fakultät für Informatik



As in classical K-Means:
The cluster center is the mean of the associated points.

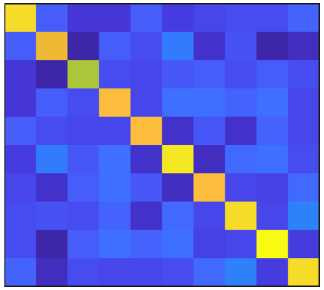
$$\boldsymbol{\mu}_{j,i} = \frac{1}{|C_{j,i}|} \sum_{\mathbf{x} \in C_{j,i}} \mathbf{x}$$

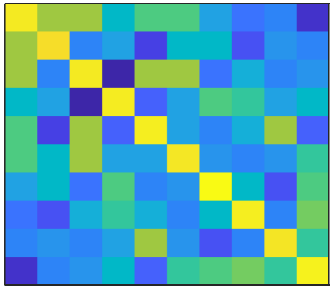
Update of the Rotation and Projection

In the case of 2 alternative clusterings in 2 subspaces, the optimal update is as follows:

- Consider the matrix $\Sigma : \Sigma_1 - \Sigma_2$
- Perform an Eigenvalue decompositon of Σ
- Sort the Eigenvectors ascendingly according to the Eigenvalues
- Update *V*: The column vectors of V are the Eigenvectors according to this sorting
- Update P_1 such that it masks the Eigenvectors corresponding to negative Eigenvalues
- Update P_2 such that it masks the remaining (positive) Eigenvectors

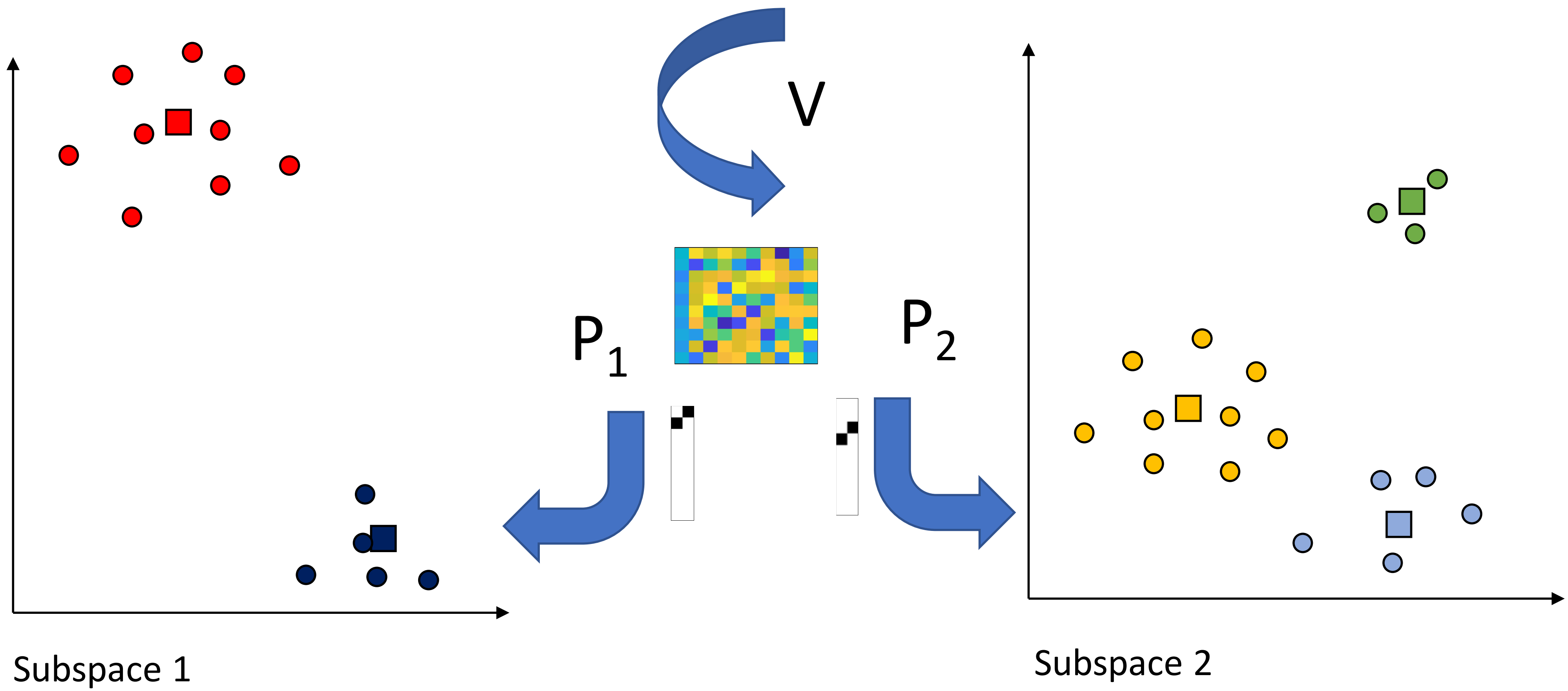
$$\Sigma_j := \sum_{i=1}^{k_j} \sum_{\mathbf{x} \in C_{j,i}} \left(\mathbf{x} - \boldsymbol{\mu}_{j,i}\right) \left(\mathbf{x} - \boldsymbol{\mu}_{j,i}\right)^{\mathrm{T}}$$ Sum of k_j scatter matrices of the clustering in Subspace *j*





Intuition: Minimize the scatter in each subspace, maximize the difference between scatters

universität wien
Fakultät für Informatik



Proof - Sketch

Uses the Trace-Trick:

We can re-write our cost function as a trace minimization problem to obtain an Eigenvalue problem

$$\mathcal{F} = \left[\sum_{i=1}^{k_1} \sum_{x \in C_{1,i}} \left\| P_1^{\mathsf{T}} V^{\mathsf{T}} \mathbf{x} - P_1^{\mathsf{T}} V^{\mathsf{T}} \boldsymbol{\mu}_{1,i} \right\|^2 \right] + \left[\sum_{i=1}^{k_2} \sum_{x \in C_{2,i}} \left\| P_2^{\mathsf{T}} V^{\mathsf{T}} \mathbf{x} - P_2^{\mathsf{T}} V^{\mathsf{T}} \boldsymbol{\mu}_{2,i} \right\|^2 \right]$$

$$= \mathrm{Tr}\left(P_1 P_1^{\mathsf{T}} V^{\mathsf{T}} [\Sigma_1 - \Sigma_2] V\right) + \mathrm{Tr}\left(V^{\mathsf{T}} \Sigma_2 V\right)$$

- A scalar is a 1x1 matrix
- Equal to its trace
- Characteristics of P_1 and P_2:

unique assignment of dimensions

Proof - Sketch

universität wien
Fakultät für Informatik

Uses the Trace-Trick:

We can re-write our cost function as a trace minimization problem to obtain an Eigenvalue problem

$$\mathcal{F} = \left[\sum_{i=1}^{k_1} \sum_{x \in C_{1,i}} \left\| P_1^{\mathsf{T}} V^{\mathsf{T}} \mathbf{x} - P_1^{\mathsf{T}} V^{\mathsf{T}} \boldsymbol{\mu}_{1,i} \right\|^2 \right] + \left[\sum_{i=1}^{k_2} \sum_{x \in C_{2,i}} \left\| P_2^{\mathsf{T}} V^{\mathsf{T}} \mathbf{x} - P_2^{\mathsf{T}} V^{\mathsf{T}} \boldsymbol{\mu}_{2,i} \right\|^2 \right]$$

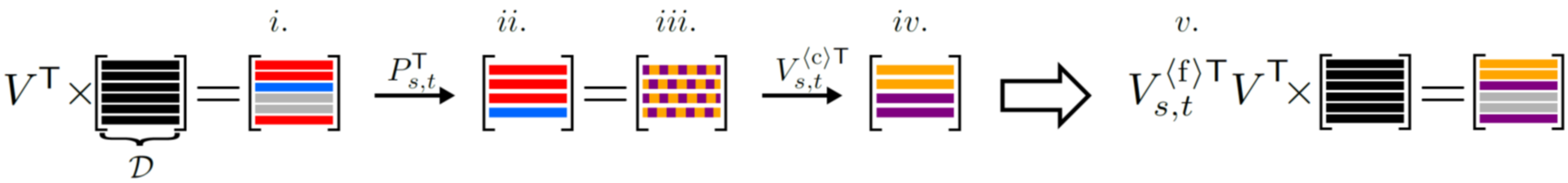
$$= \operatorname{Tr}\left(P_1 P_1^{\mathsf{T}} V^{\mathsf{T}} \left[\Sigma_1 - \Sigma_2\right] V\right) + \operatorname{Tr}\left(V^{\mathsf{T}} \Sigma_2 V\right)$$

- A scalar is a 1x1 matrix
- Equal to its trace
- Characteristics of P_1 and P_2: unique assignment of dimensions

Automatic selection of dimensionality:

We minimize our objective function by assigning Eigenvectors with negative Eigenvalues to S_1 and the others to S_2

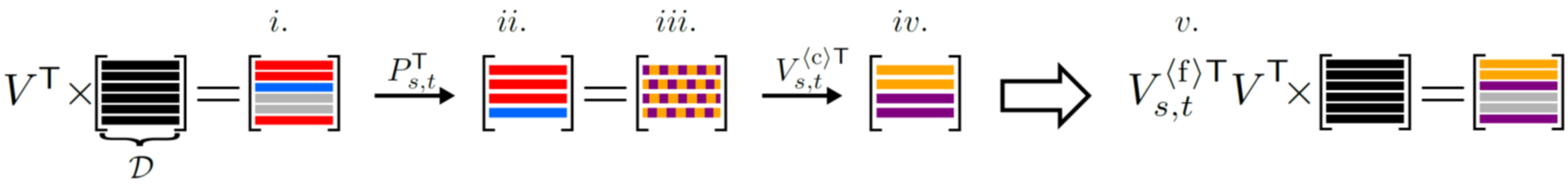
More than 2 Clusterings – Pairwise updates



Consider all pairs of subspaces:
i) -> ii) Project pair to the joint space,
ii) -> iii) optimize rotation in the projected space
iii -> iv) determine best dimensionality
iv -> v) propagate these changes to the full dimensional space

More than 2 Clusterings – Pairwise updates

universität wien
Fakultät für Informatik



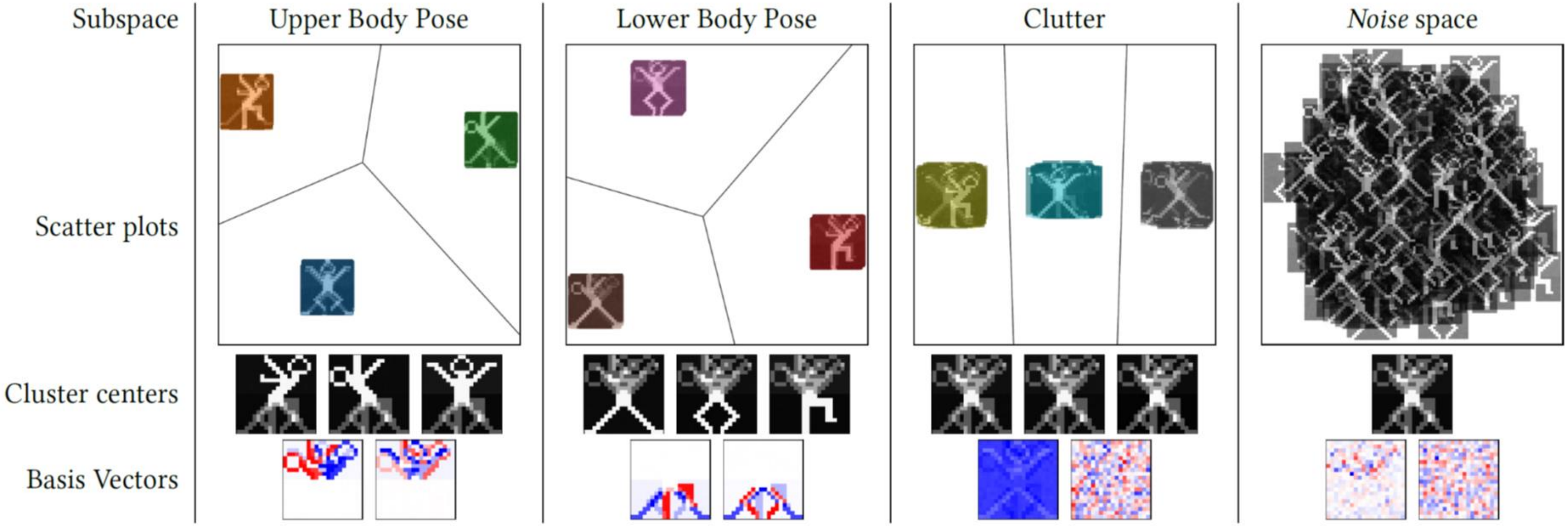
Consider all pairs of subspaces:
i) -> ii) Project pair to the joint space,
ii) -> iii) optimize rotation in the projected space
iii -> iv) determine best dimensionality
iv -> v) propagate these changes to the full dimensional space

Runtime Complexity:

as classical K-Means linear in the number of iterations, the number of clusters and data objects;
Additional cubic complexity in the dimensionality for Eigenvalue decomposition

NR-K-Means: Experiments

UCI Stickfigures dataset, 900 objects, 400 dimensions



Outline

universität wien
Fakultät für Informatik

1. Introduction
2. Alternative Clustering
3. Autoencoders
4. Deep Embedded Non-Redundant Clustering
5. Application to Archeology
6. Conclusion and Outlook

Moving to higher Dimensions by Deep Learning

universität wien
Fakultät für Informatik

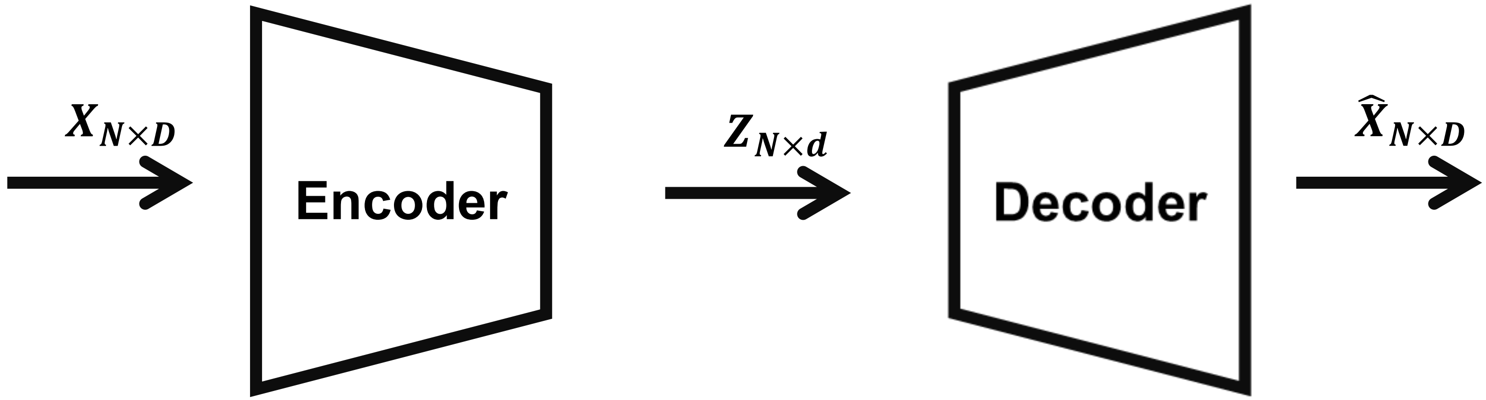
- Successful for image, text, video, audio …
  - Structured data
  - High data volume
- Automated feature extraction (Representation Learning)
  - Useful for supervised and unsupervised learning
  - Feature engineering requires domain knowledge
- Easy to parallelize
  - GPU friendly
  - Works on large amount of data

Autoencoders

- Learning is done via self-supervision – requires no labels

- The prediction (output) is a reconstruction of the input data
- Goal: Low dimensional representation (embedding) of input data

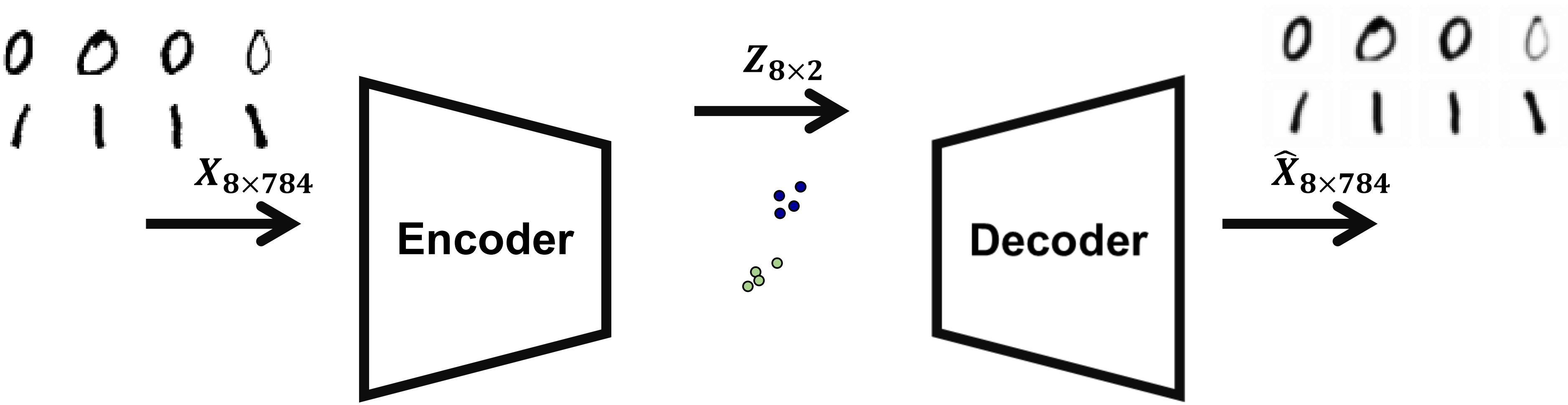
  Sketch of an autoencoder architecture:



Autoencoders – Toy Example

- Learning is done via self-supervision – requires no labels

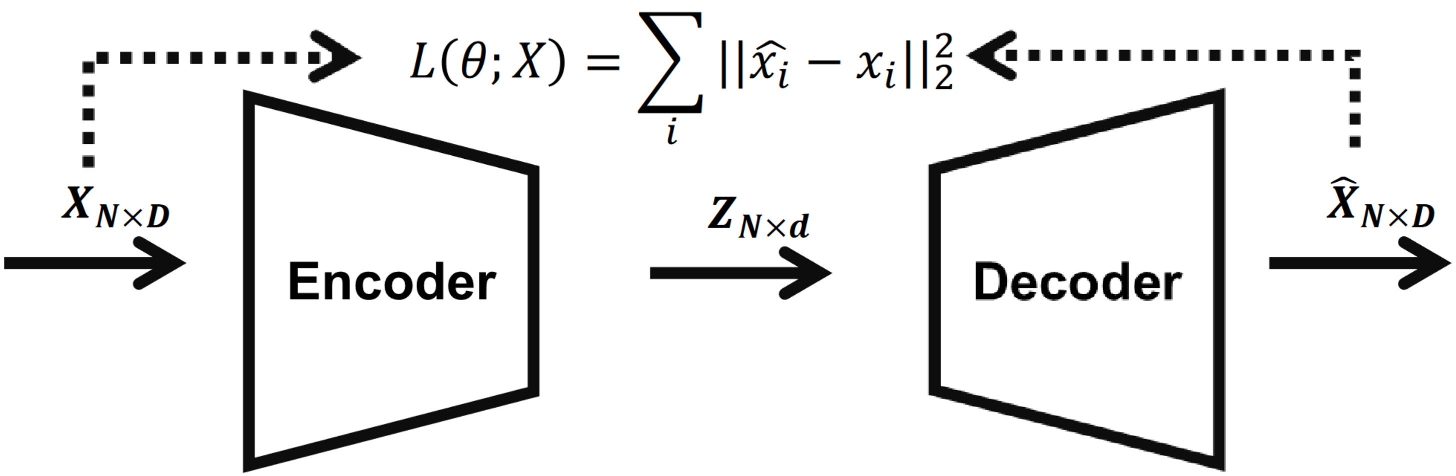
- The prediction (output) is a reconstruction of the input data
- Goal: Low dimensional representation (embedding) of input data



Digits from: https://de.wikipedia.org/wiki/MNIST-Datenbank#/media/Datei:MnistExamples.png

Loss Function

- Compares the reconstruction $\hat{x}$ with the input x
- Quantifies the reconstruction loss which we want to minimize
- Common choices for loss functions:
  - For binary inputs: Cross Entropy
  - For real valued inputs: Sum of Squared Differences



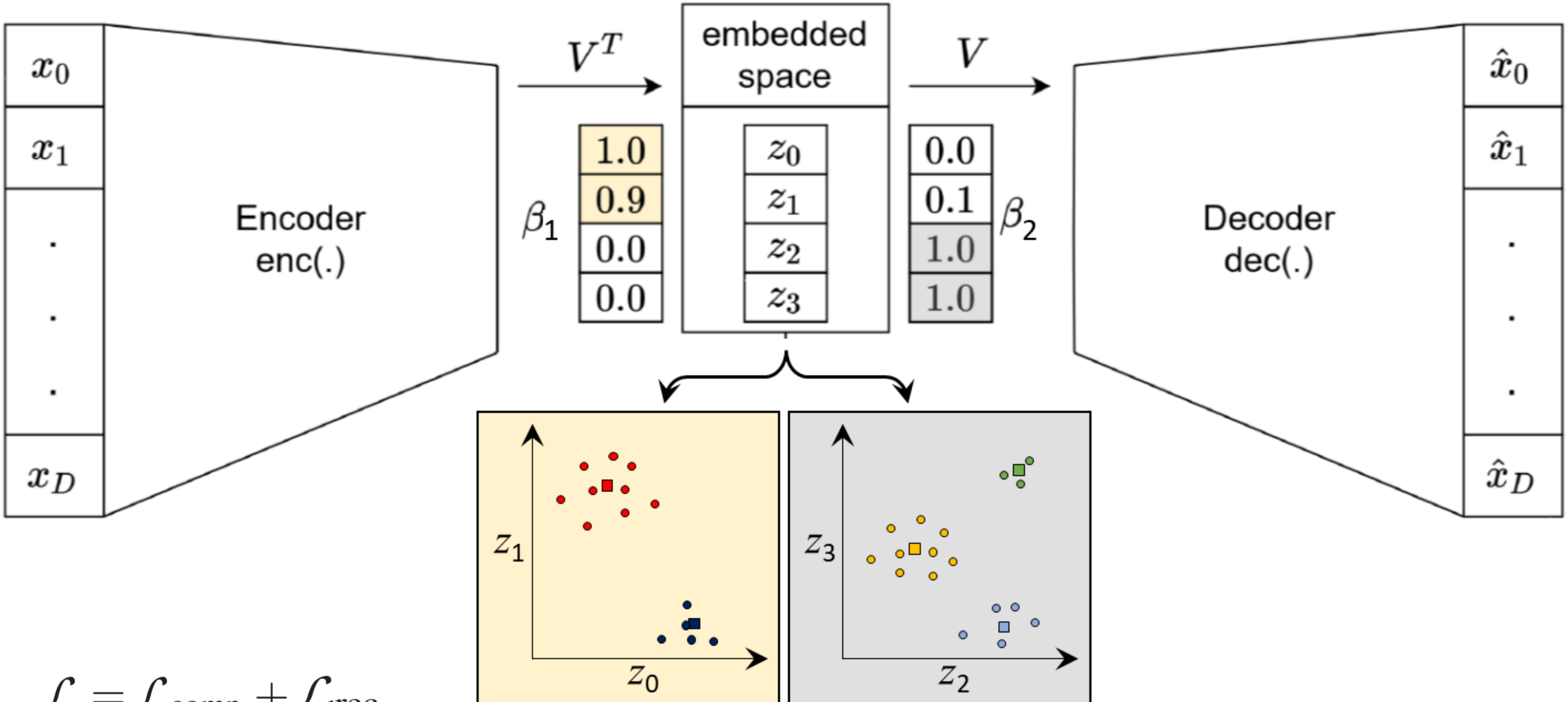
Where Θ are all learnable parameters of the autoencoder

Outline

universität wien
Fakultät für Informatik

1. Introduction
2. Alternative Clustering
3. Autoencoders
4. Deep Embedded Non-Redundant Clustering
5. Application to Archeology
6. Conclusion and Outlook

Architecture of ENRC (Deep Embedded Non-redundant Clustering)



$$\mathcal{L} = \mathcal{L}_{\text{comp}} + \mathcal{L}_{\text{rec}}$$

$$\mathcal{L}_{\text{comp}} = \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\frac{1}{|C_{s,k}|}\sum_{z\in C_{s,k}} ||V^{\mathsf{T}}z - \mu_{s,k}||^2_{\beta_s}$$

$$\mathcal{L}_{\text{rec}} = ||\mathbf{x} - \text{dec}(VV^{\mathsf{T}}\text{enc}(\mathbf{x}))||_2^2$$

Algorithm ENRC – Initialization Phase

Init embeddings and cluster centers:

- pre-train autoencoder,
- init V as random orthogonal matrix
- Initial strong assignments of dimensions to clusterings ($\beta = 0.9$),
- K-Means in subspaces to get initial μ

Keep the autoencoder parameters fixed and optimize V, β, μ

Algorithm ENRC – Clustering Phase

universität wien
Fakultät für Informatik

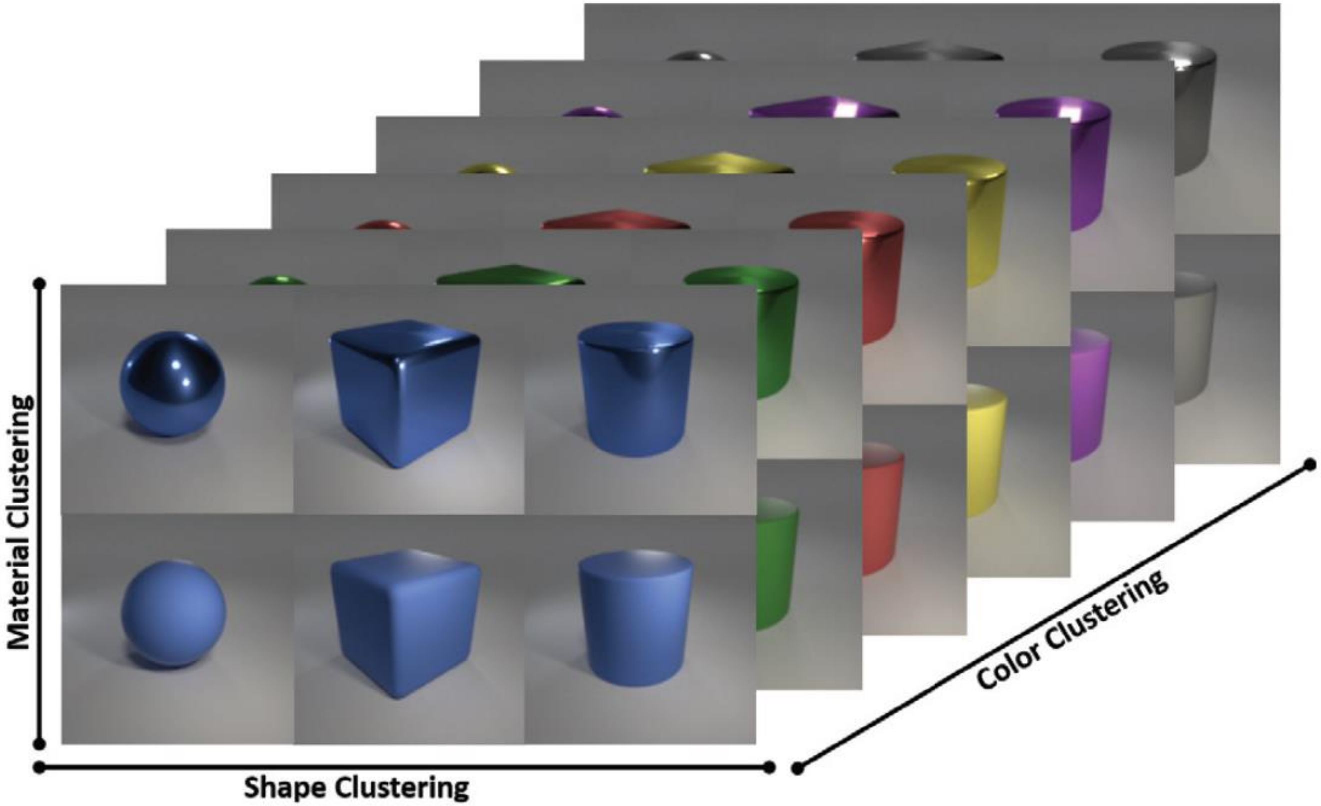
Optimize all parameters by mini-batch training:

- The cluster assignments and embeddings of all objects,
- The cluster centers μ
- The dimension weights β
- The rotation matrix V

$$\mathcal{L} = \mathcal{L}_{\text{comp}} + \lambda \mathcal{L}_{\text{rec}}$$

$$\mathcal{L}_{\text{rec}} = ||\mathbf{x} - \text{dec}(VV^{\top}\text{enc}(\mathbf{x}))||_2^2, \qquad \mathcal{L}_{\text{comp}} = \frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K_s}\frac{1}{|C_{s,k}|}\sum_{z \in C_{s,k}} ||V^{\top}z - \mu_{s,k}||_{\beta_s}^2.$$

ENRC - Experiments

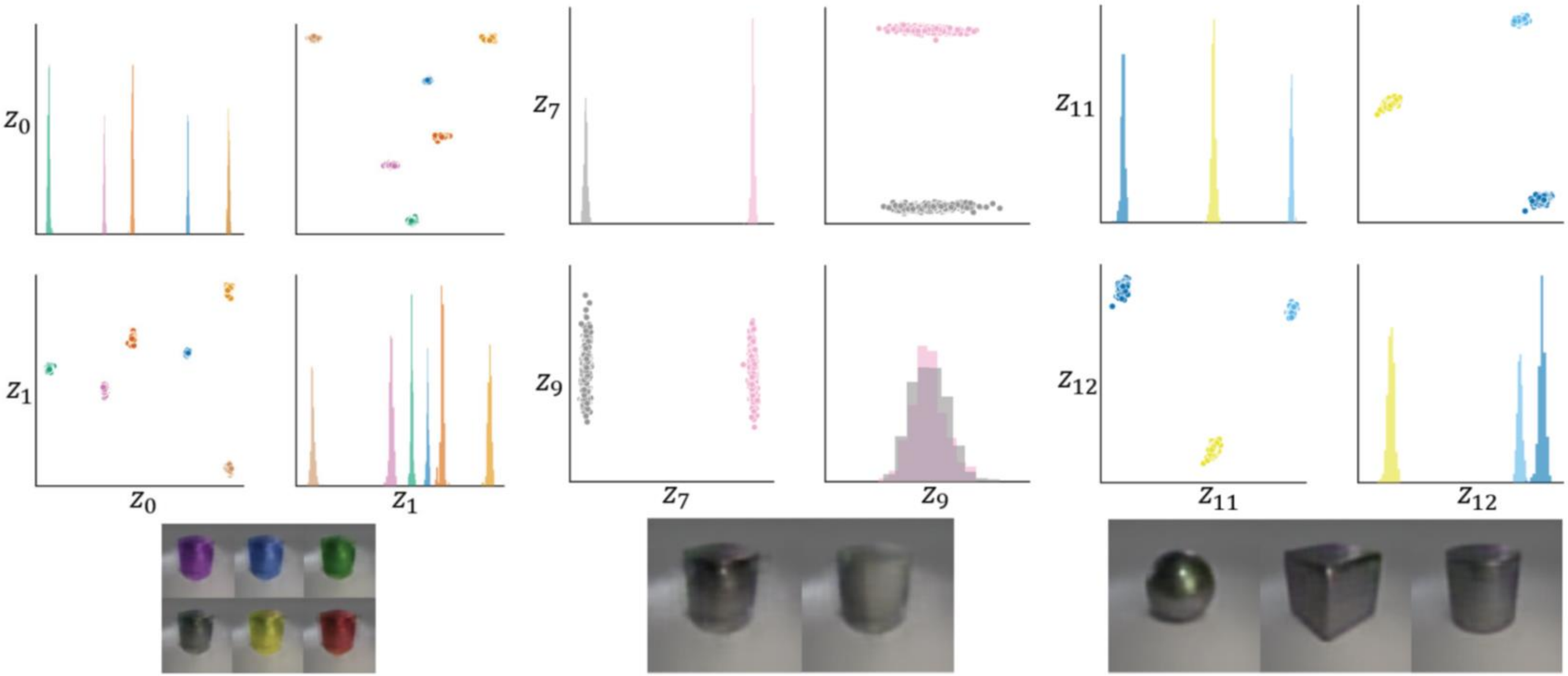


NR-Objects data

16,384 dimensions
10,000 objects

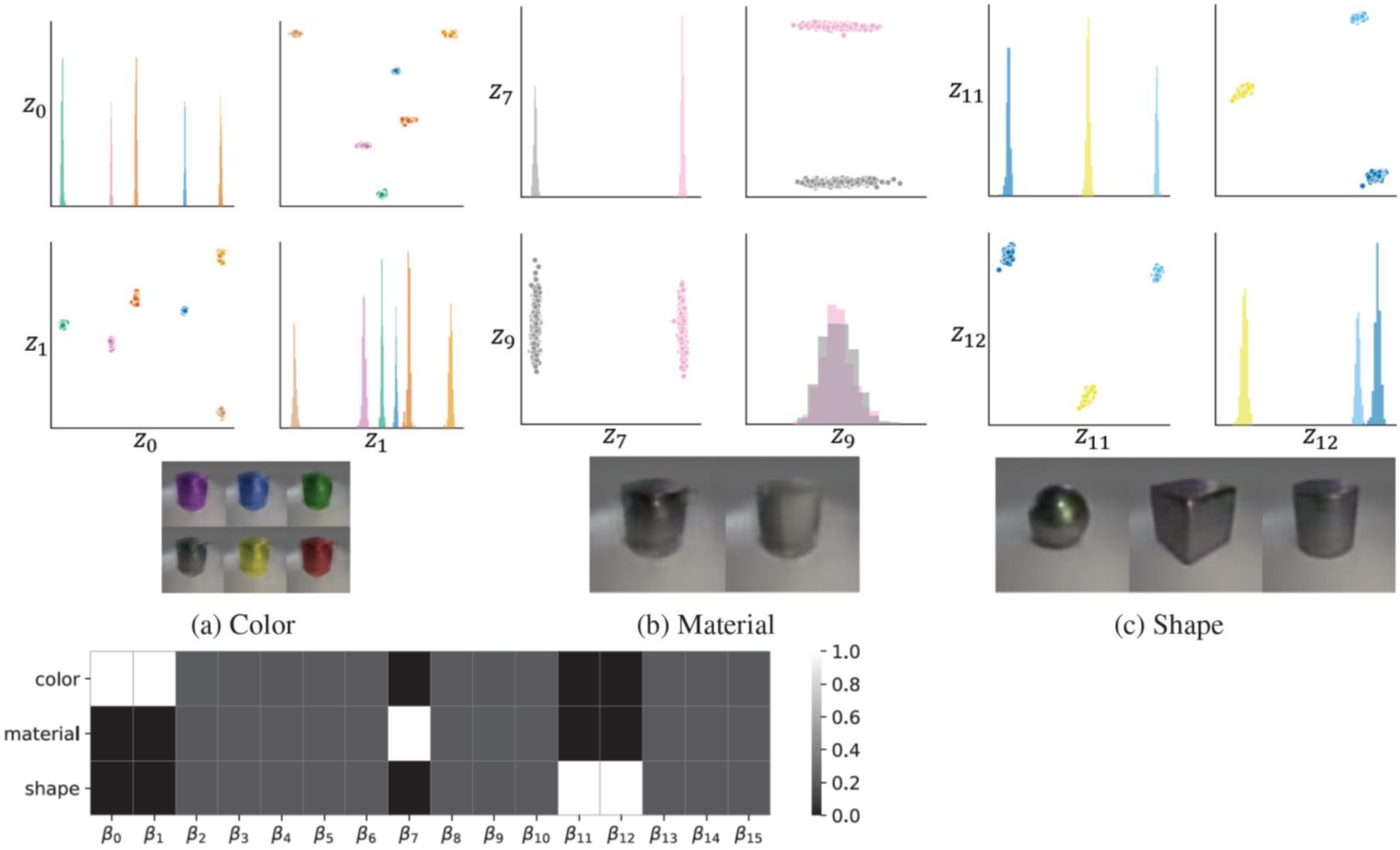
https://github.com/facebookresearch/clevr-dataset-gen

ENRC - Experiments



(a) Color (b) Material (c) Shape

ENRC - Experiments



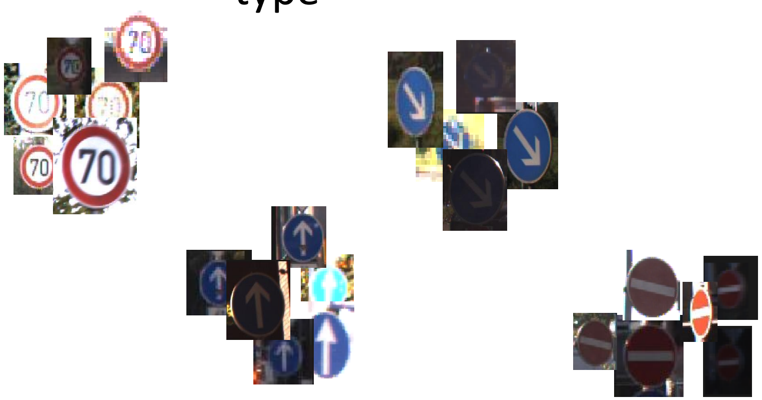
ENRC - Experiments

Data Sets	Clustering	ENRC	Orth1	Orth2	mSC	Nr-Kmeans	ISAAC
NR-Objects	color	**1.00 ±0.00**	0.70 ±0.09	0.73 ±0.06	0.35 ±0.05	0.92 ±0.09	0.15 ±0.06
	material	**1.00 ±0.00**	0.46 ±0.16	0.11 ±0.12	0.03 ±0.07	0.95 ±0.14	0.53 ±0.08
	shape	**1.00 ±0.00**	0.39 ±0.20	0.20 ±0.08	0.03 ±0.03	0.92 ±0.16	0.60 ±0.07
GTSRB	type	**0.74 ±0.01**	0.57 ±0.07	0.73 ±0.15	0.04 ±0.04	0.72 ±0.01	0.60 ±0.07
	color	**0.67 ±0.00**	0.59 ±0.02	0.63 ±0.03	0.04 ±0.06	0.65 ±0.01	0.59 ±0.04
Stickfigures	upper	**1.00 ±0.00**	0.79 ±0.21	0.00 ±0.00	0.33 ±0.20	**1.00 ±0.00**	0.37 ±0.05
	lower	**1.00 ±0.00**	0.77 ±0.24	0.00 ±0.00	0.30 ±0.17	**1.00 ±0.00**	0.39 ±0.08
C-MNIST	left	**0.83 ±0.04**	0.33 ±0.02	0.35 ±0.03	0.07 ±0.02*	0.69 ±0.03	0.29 ±0.13*
	right	**0.82 ±0.01**	0.40 ±0.03	0.41 ±0.04	0.06 ±0.02*	0.70 ±0.03	0.19 ±0.13*

ENRC - Experiments

Data Sets	Clustering	ENRC	Orth1	Orth2	mSC	Nr-Kmeans	ISAAC
NR-Objects	color	**1.00 ±0.00**	0.70 ±0.09	0.73 ±0.06	0.35 ±0.05	0.92 ±0.09	0.15 ±0.06
	material	**1.00 ±0.00**	0.46 ±0.16	0.11 ±0.12	0.03 ±0.07	0.95 ±0.14	0.53 ±0.08
	shape	**1.00 ±0.00**	0.39 ±0.20	0.20 ±0.08	0.03 ±0.03	0.92 ±0.16	0.60 ±0.07
GTSRB	type	**0.74 ±0.01**	0.57 ±0.07	0.73 ±0.15	0.04 ±0.04	0.72 ±0.01	0.60 ±0.07
	color	**0.67 ±0.00**	0.59 ±0.02	0.63 ±0.03	0.04 ±0.06	0.65 ±0.01	0.59 ±0.04
Stickfigures	upper	**1.00 ±0.00**	0.79 ±0.21	0.00 ±0.00	0.33 ±0.20	**1.00 ±0.00**	0.37 ±0.05
	lower	**1.00 ±0.00**	0.77 ±0.24	0.00 ±0.00	0.30 ±0.17	**1.00 ±0.00**	0.39 ±0.08
C-MNIST	left	**0.83 ±0.04**	0.33 ±0.02	0.35 ±0.03	0.07 ±0.02*	0.69 ±0.03	0.29 ±0.13*
	right	**0.82 ±0.01**	0.40 ±0.03	0.41 ±0.04	0.06 ±0.02*	0.70 ±0.03	0.19 ±0.13*

type



color

German Traffic Sign Benchmark Data

Comparison

universität wien
Fakultät für Informatik

	High-dimensional data	Interpretability	Runtime	Parameterization
K-Means	--- (up to about 10)	+++ (centroids)	+++ (milliseconds unithreaded CPU)	- (# clusters)
NR-K-Means	+ (hundreds)	++ (centroids plus eigenspaces, orthonormal rotations and projections)	++ (seconds unithreaded CPU)	-- (# clusters, # clusterings)
ENRC – the first deep alternative clustering method	+++ (several thousands)	+ (centroids, arbitrary space transformation)	--- (minutes to hours on GPU)	--- (# clusters, # clusterings, dimensionality of clustered spaces, hyperparameters of autoencoder)

Outline

universität wien
Fakultät für Informatik

1. Introduction
2. Alternative Clustering
3. Autoencoders
4. Deep Embedded Non-Redundant Clustering
5. Application to Archeology
6. Conclusion and Outlook

Alternative Clustering for Classification of Early Medieval Glass Beads

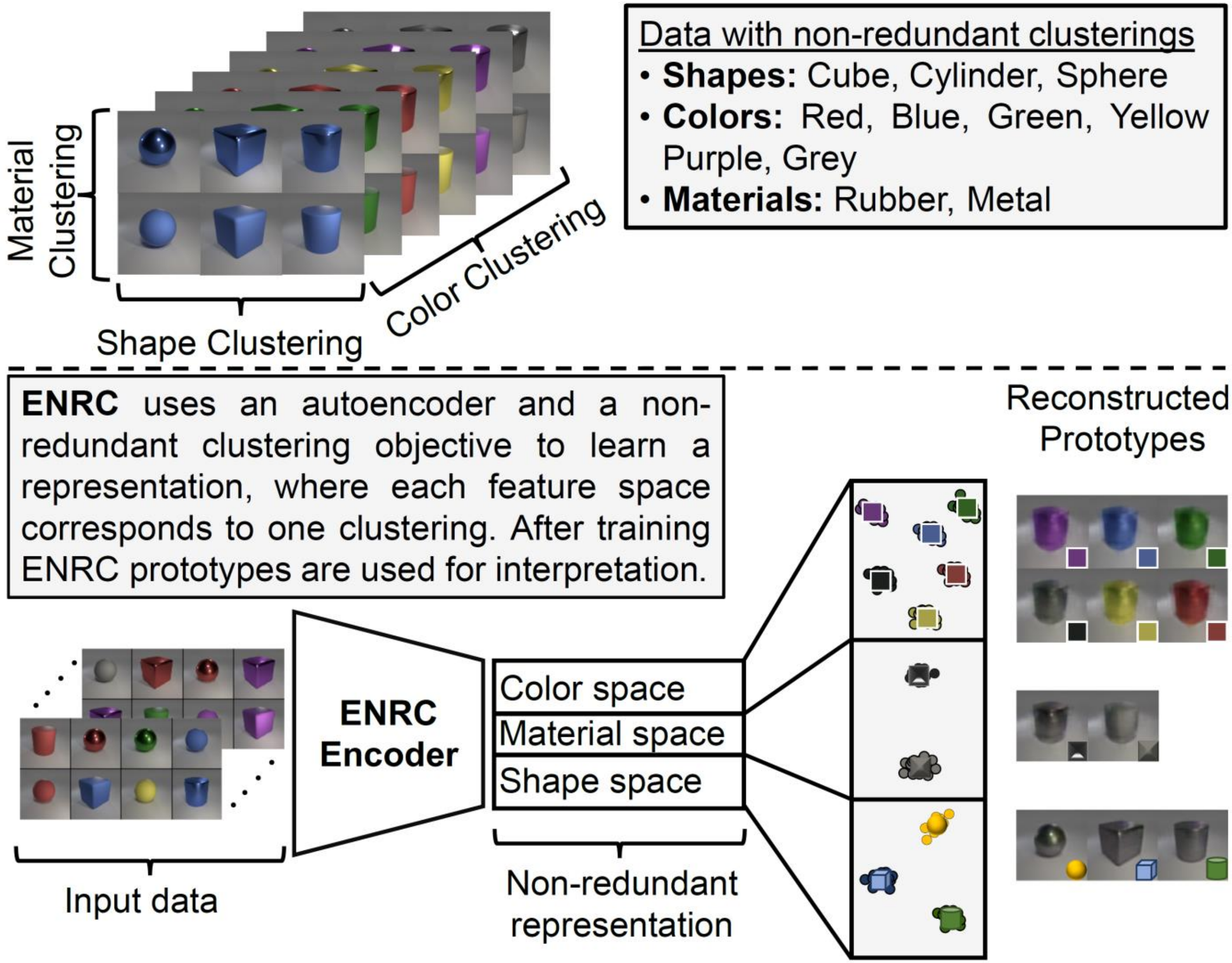
universität wien
Fakultät für Informatik

- Among the most common grave goods in the early Middle Ages
- production sites in the Middle East and Southeast Asia
- from there, most of the beads reached even the most remote areas of Europe
- The color, size, shape, production technique and decoration of the beads are diverse.
- Classification systems are often subjective, complex and mostly limited to one burial field.

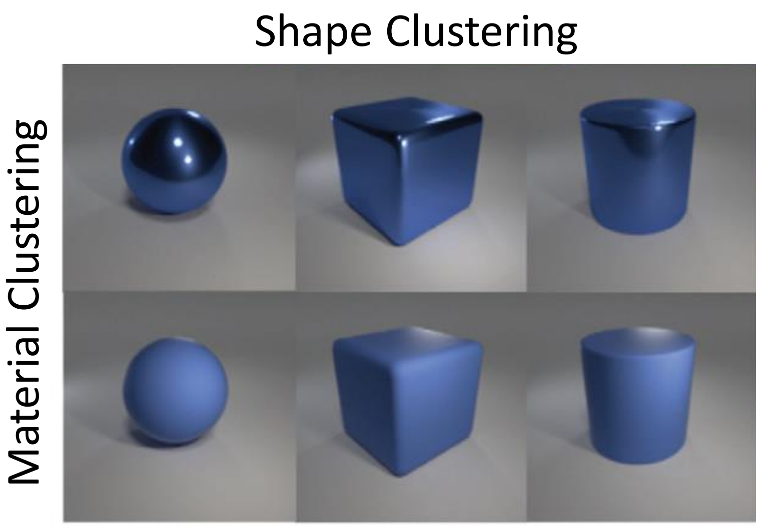
Ongoing Project: *The Glass Bead Network*
Classification of early medieval beads from Vienna-Csokorgasse using AI
Together with Bendeguz Tobias from ÖAW

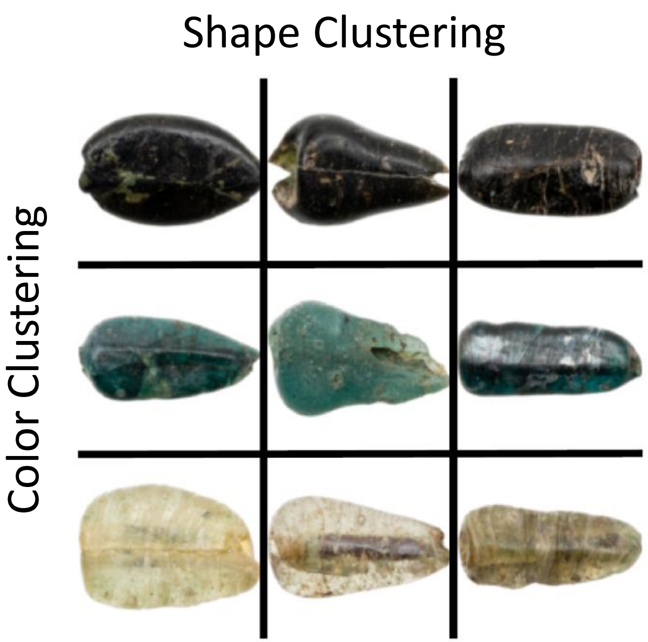
ÖAW ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

Recall: ENRC in a nutshell



Challenges when moving to the real world



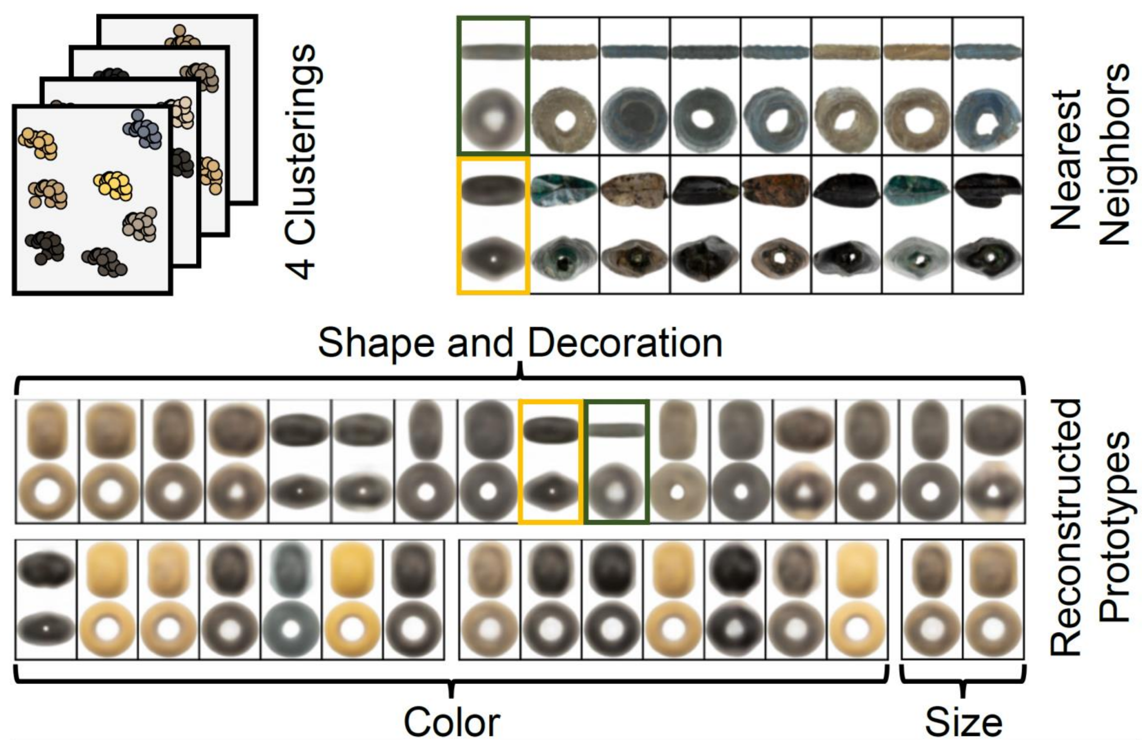


- Imbalances
- Corrupted instances due to aging and restoration
- Top and side view for each image
- Small sample size: 4669 beads
- Outliers
- Difficult parameterization
- Partially overlapping clusterings: often barrel-shaped and yellow

Analysis Pipeline

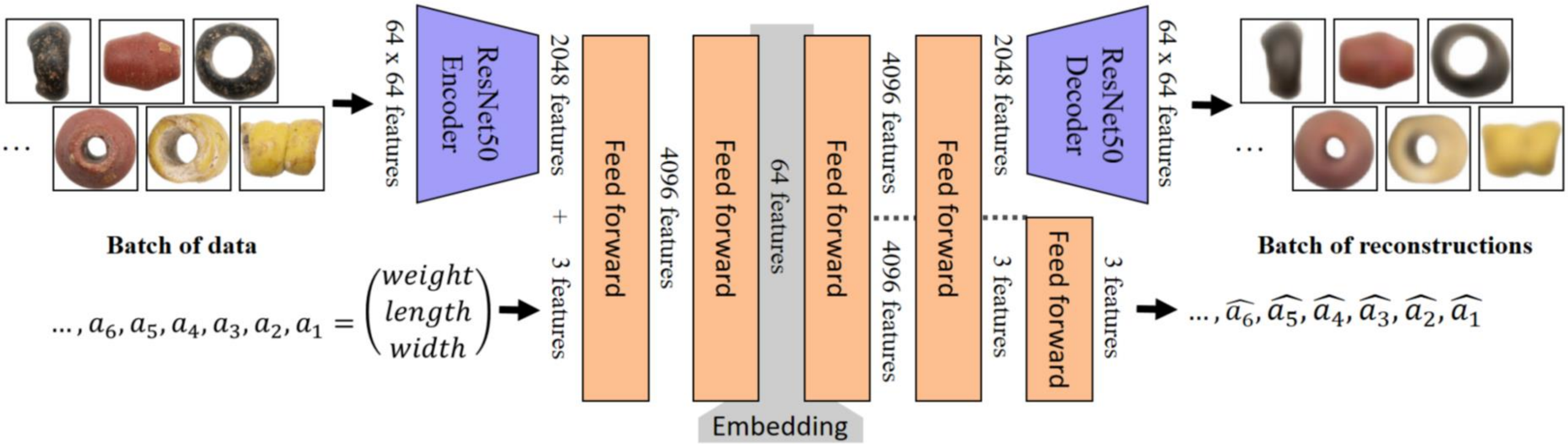
Embedded **N**on-**R**edundant **C**lustering of Early Medieval glass beads. The beads are **collected** from museums in Austria, recorded, **preprocessed** and used for **pretraining**. The number of clusterings are **estimated** with AutoNR and fine-tuned with **ENRC**. Prototypes and their nearest neighbors are used for interpretation.

Data Collection → Preprocessing → Autoencoder Pretraining → Cluster estimation → ENRC → Evaluation and Interpretation



Evaluation and Interpretation

Autoencoder Pretraining



Mixed convolutional autoencoder based on ResNet.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition. CVPR 2016: 770-778

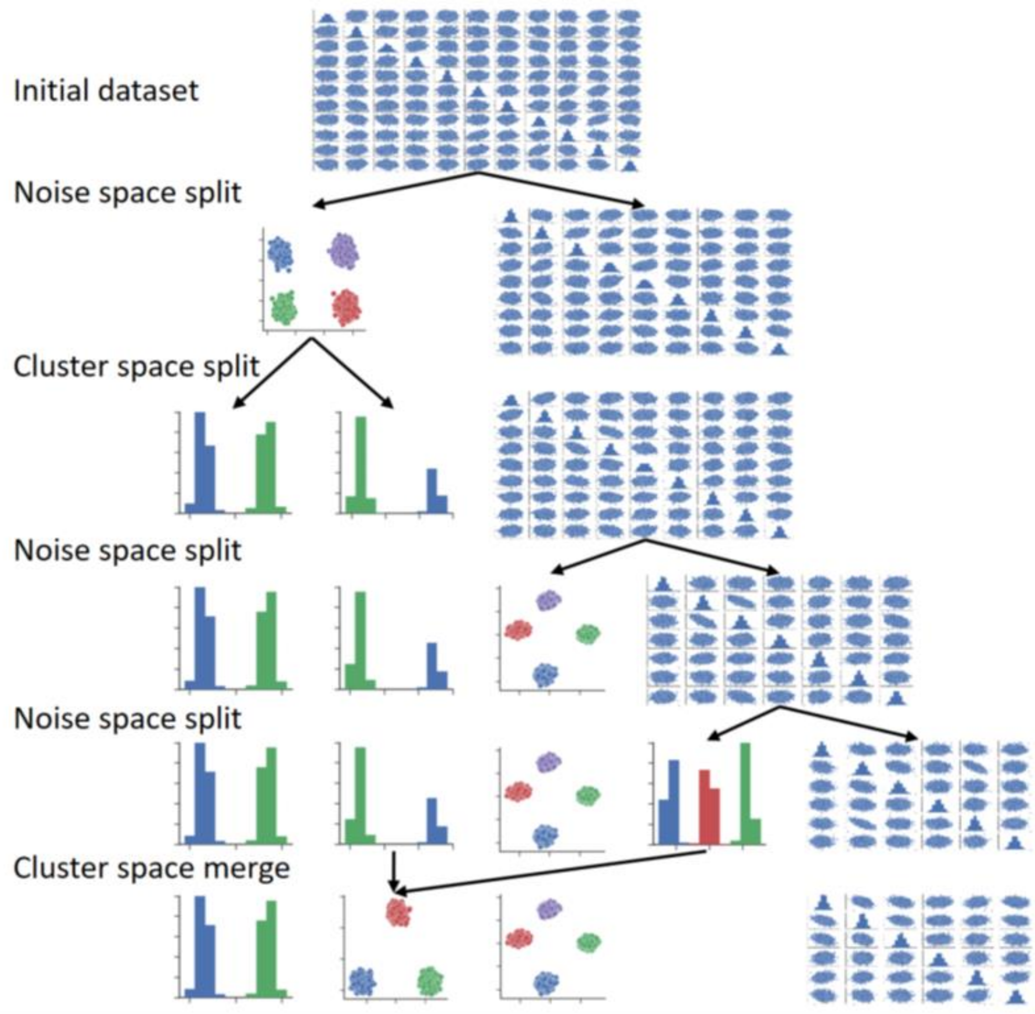
Information-theoretic Parameter Estimation

Auto-NR: Greedy algorithm relying on the Minimum Description Length Principle **to find suitable parameters for NR-K-means.**

In each step:
Choose the operation that best improves the coding costs of the data given the cluster model.

Also supports **identification of outliers.**

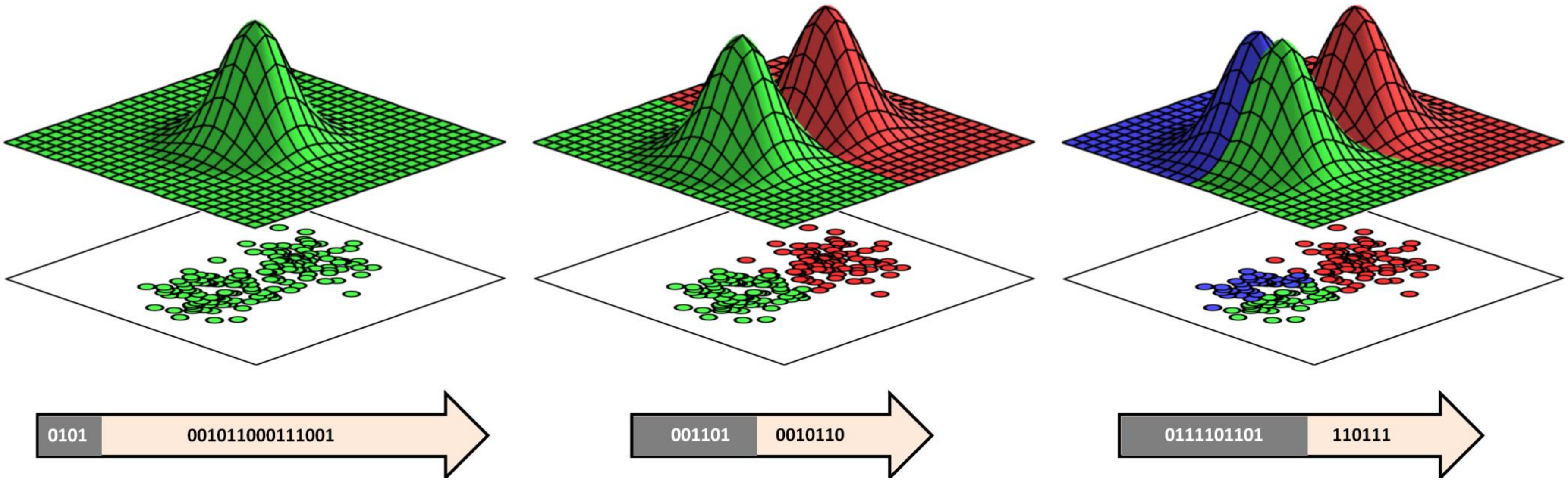
Selection of subspace dimensionality: Experimentally.
Similar results for 64 and 128D, 32D seems not enough.



Collin Leiber, Dominik Mautz, Claudia Plant, Christian Böhm:Automatic Parameter Selection for Non-Redundant Clustering. SDM 2022: 226-234

Information-theoretic Parameter Estimation

Selection of the number of clusters by data compression.



ENRC with Application-specific Augmentations

universität wien
Fakultät für Informatik

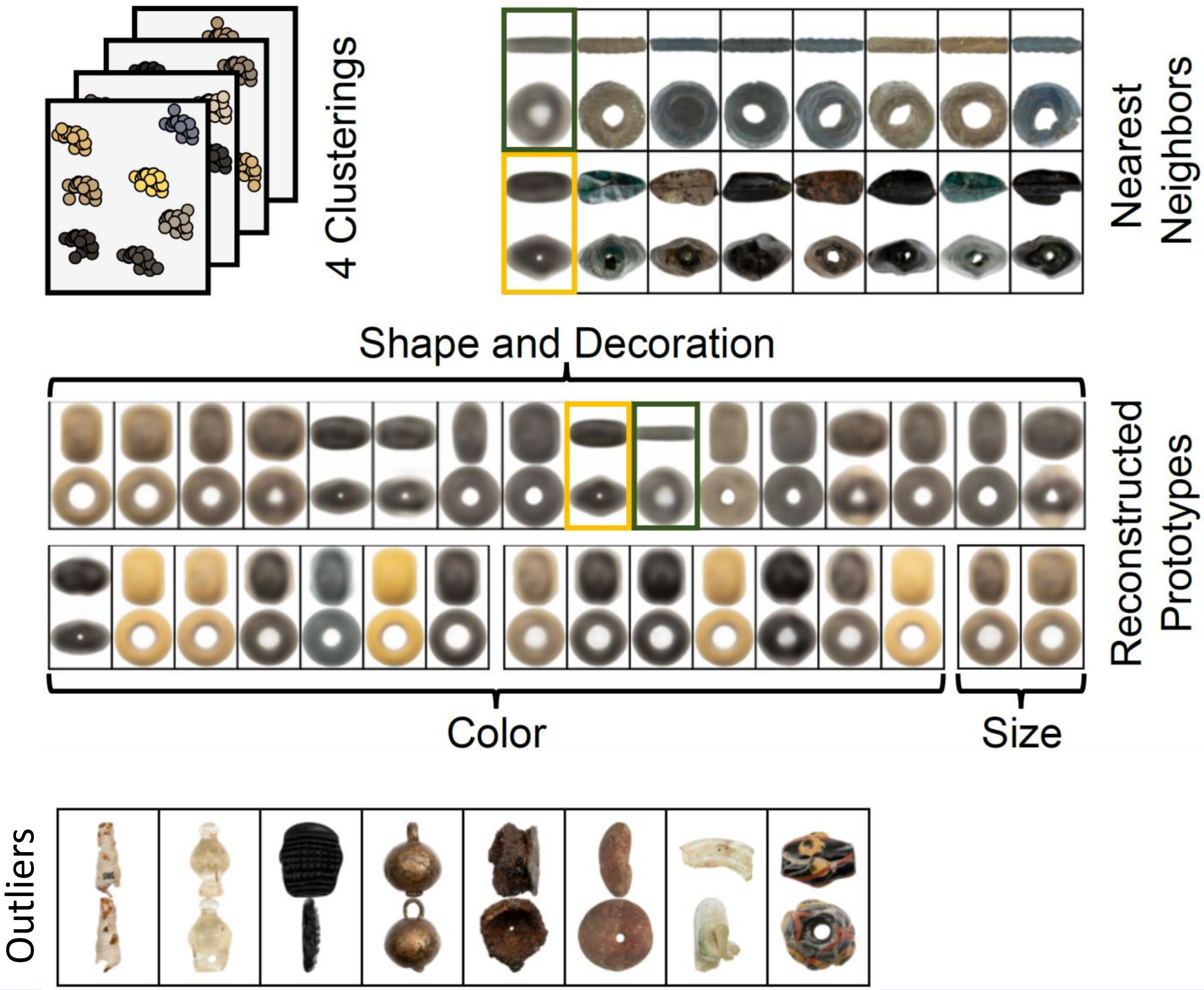
$$\sum_{j=1}^{J}\sum_{k=1}^{K_j}\sum_{z_{\text{top}},z_{\text{side}}\in C_{j,k}} ||V^T z_{\text{top}} - V^T \mu_{j,k}||^2_{\beta_j} + ||V^T z_{\text{side}} - V^T \mu_{j,k}||^2_{\beta_j}$$

z_{top} Encoded top view with augmentations to achieve invariance against horizontal and vertical flipping,
Slight rotations and transformations; cropping to account for missing parts; color augmentations to cope with
imbalances

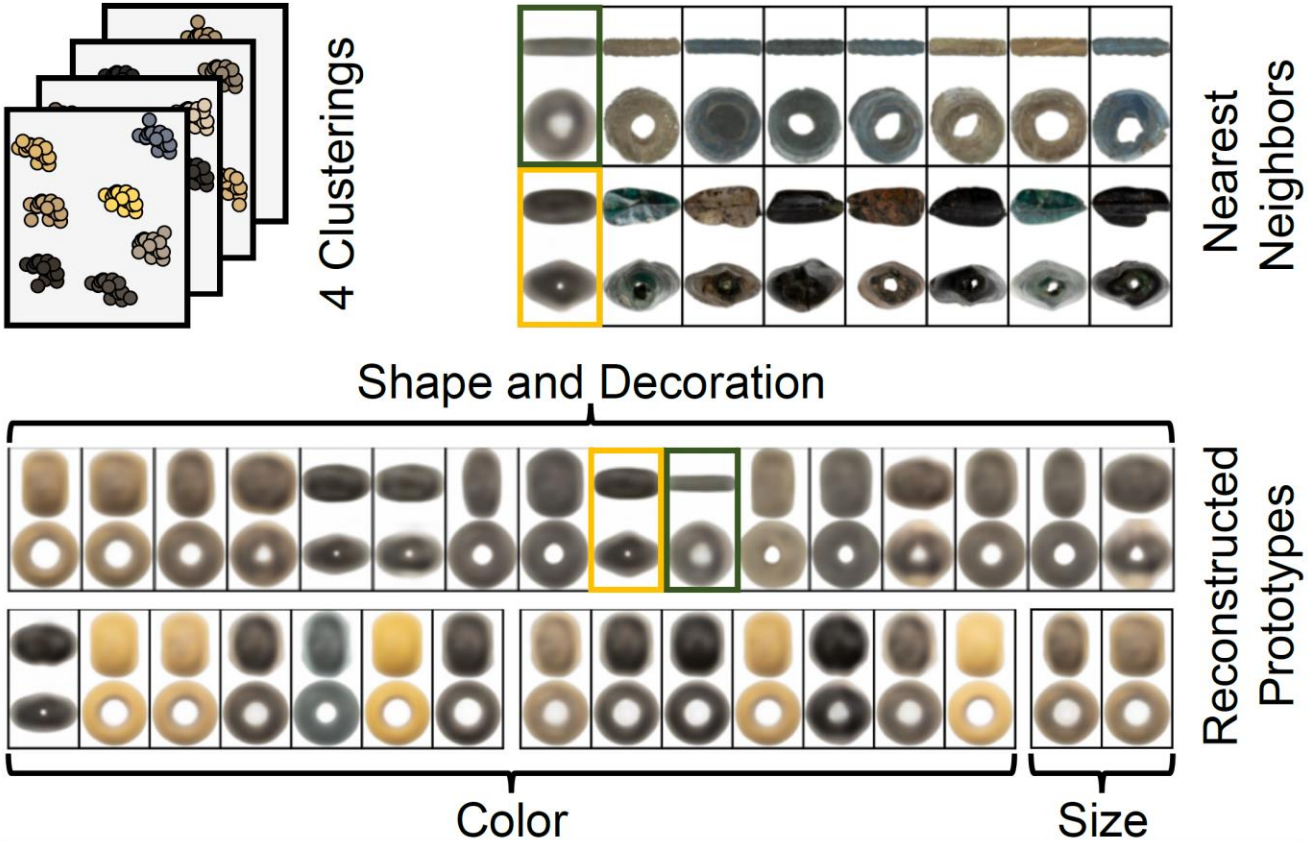
Original images

Augmented images

Results


4 Clusterings
Nearest Neighbors
Shape and Decoration
Reconstructed Prototypes
Color
Size
Outliers

Results: A Fingerprint of Vienna-Csokorgasse



Summary

- To the best of our knowledge the first application of alternative clustering to ancient glass beads
- The results summarize the findings of a burial site and support objective comparison of the findings of different sites – like a fingerprint of a burial site
- Lots of interesting but incomplete further information that we currently do not use, e.g. beads belonging to one necklace, beads found in one grave, beads found at a certain depth

Outline

universität wien
Fakultät für Informatik

1. Introduction
2. Alternative Clustering
3. Autoencoders
4. Deep Embedded Non-Redundant Clustering
5. Application to Archeology
6. Conclusion and Outlook

Comparison

universität wien
Fakultät für Informatik

	High-dimensional data	Interpretability	Runtime	Parameterization
K-means	--- (up to about 10)	+++ (centroids)	+++ (milliseconds unithreaded CPU)	- (# clusters)
NR-K-means	+ (hundreds)	++ (centroids plus eigenspaces, orthonormal rotations and projections)	++ (seconds unithreaded CPU)	-- (# clusters, # clusterings)
ENRC	+++ (serveral thousands)	+ (centroids, arbitrary space transformation)	--- (minutes to hours on GPU)	--- (# clusters, # clusterings, dimensionality of clustered spaces, hyperparameters of autoencoder)

Looking at this from a more general perspective…

universität wien
Fakultät für Informatik

	High-dimensional data	Interpretability	Runtime	Parameterization
Traditional clustering algorithms, e.g. K-means (1950 and older)	---	+++	+++	-
Subspace and spectral methods, e.g., NR-K-means (starting in the 1990ies)	+	++	++	--
Deep clustering methods, e.g., ENRC (popular since 2010)	+++	+	---	---

…hybrid methods might be the future.

	High-dimensional data	Interpretability	Runtime	Parameterization
Traditional clustering algorithms	---	+++	+++	-
Subspace and spectral methods	+	++	++	--
Deep clustering methods	+++	+	---	---
Hybrid methods	+++ expressiveness where needed?	++ interpretable where possible?	+ spend effort where needed?	-- partly automatic?

We need a cost model/objective function for hybrid methods

that supports answering the questions:

- How much model complexity/expressiveness do we need to cluster our data?
- How to trade-off the gain in expressiveness by deep clustering methods with the excessive runtime and energy demand?

We need a cost model/objective function for hybrid methods

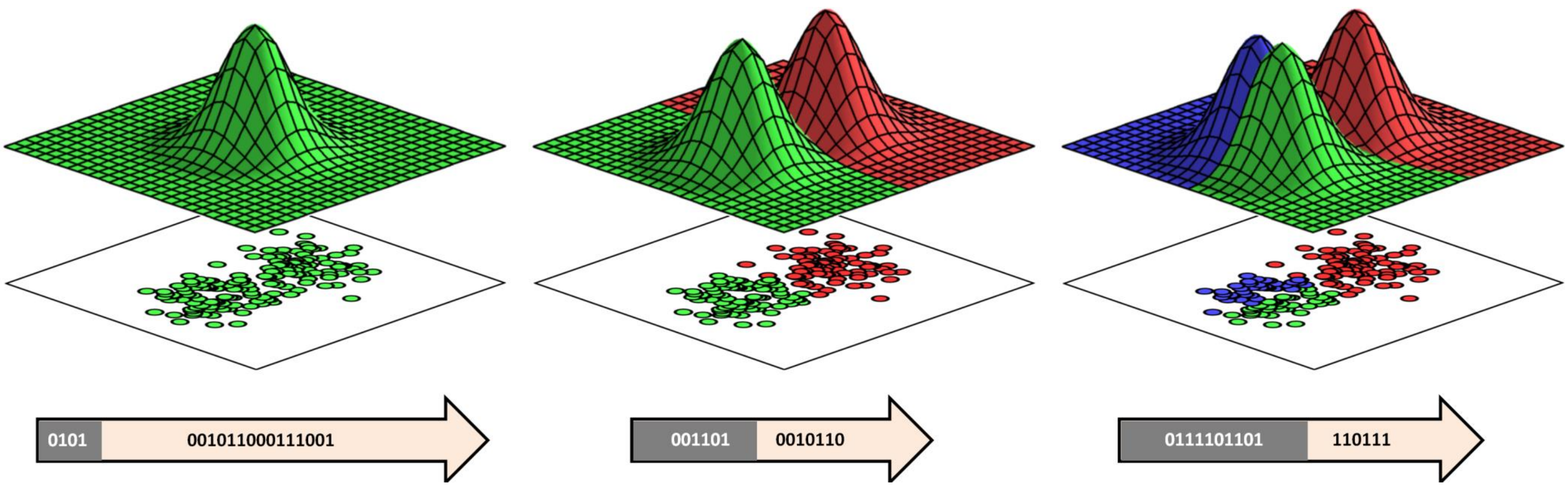
that supports answering the questions:

- How much model complexity/expressiveness do we need to cluster our data?
- How to trade-off the gain in expressiveness by deep clustering methods with the excessive runtime and energy demand?

Data Sets	Clustering	ENRC	Orth1	Orth2	mSC	Nr-Kmeans	ISAAC
NR-Objects	color	**1.00** $\pm$**0.00**	0.70 $\pm$0.09	0.73 $\pm$0.06	0.35 $\pm$0.05	0.92 $\pm$0.09	0.15 $\pm$0.06
	material	**1.00** $\pm$**0.00**	0.46 $\pm$0.16	0.11 $\pm$0.12	0.03 $\pm$0.07	0.95 $\pm$0.14	0.53 $\pm$0.08
	shape	**1.00** $\pm$**0.00**	0.39 $\pm$0.20	0.20 $\pm$0.08	0.03 $\pm$0.03	0.92 $\pm$0.16	0.60 $\pm$0.07
GTSRB	type	**0.74** $\pm$**0.01**	0.57 $\pm$0.07	0.73 $\pm$0.15	0.04 $\pm$0.04	0.72 $\pm$0.01	0.60 $\pm$0.07
	color	**0.67** $\pm$**0.00**	0.59 $\pm$0.02	0.63 $\pm$0.03	0.04 $\pm$0.06	0.65 $\pm$0.01	0.59 $\pm$0.04
Stickfigures	upper	**1.00** $\pm$**0.00**	0.79 $\pm$0.21	0.00 $\pm$0.00	0.33 $\pm$0.20	**1.00** $\pm$**0.00**	0.37 $\pm$0.05
	lower	**1.00** $\pm$**0.00**	0.77 $\pm$0.24	0.00 $\pm$0.00	0.30 $\pm$0.17	**1.00** $\pm$**0.00**	0.39 $\pm$0.08
C-MNIST	left	**0.83** $\pm$**0.04**	0.33 $\pm$0.02	0.35 $\pm$0.03	0.07 $\pm$0.02*	0.69 $\pm$0.03	0.29 $\pm$0.13*
	right	**0.82** $\pm$**0.01**	0.40 $\pm$0.03	0.41 $\pm$0.04	0.06 $\pm$0.02*	0.70 $\pm$0.03	0.19 $\pm$0.13*

We need a cost model/objective function for hybrid methods

For traditional and subspace clustering methods, data compression works, **but how to deal with huge parameter spaces**? (ResNet50: about 100 millions of trainable parameters to model about 5000 glass beads)



And how to tackle energy efficiency?

Some solved and a lot more open problems – so the journey will go on ☺

universität wien
Fakultät für Informatik

Dr. Dominik Mautz
PhD 2022 (LMU)

Lukas Miklautz
PhD thesis submitted (UniVie)

Dr. Bendeguz Tobias
Glass Beads Project (ÖAW)

Prof. Wei Ye
PhD 2018 (LMU)
Now TT-Prof. at Tongji University, Shanghai

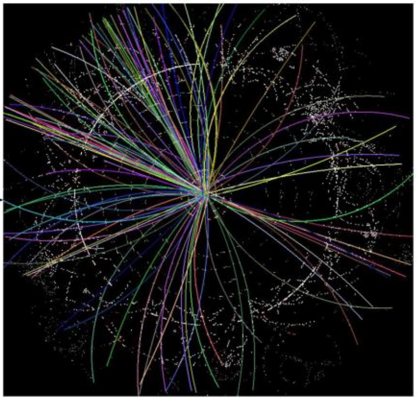
Collin Leiber, LMU

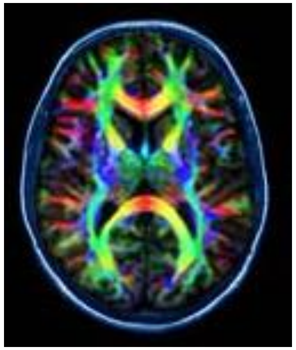
Lena Bauer, UniVie

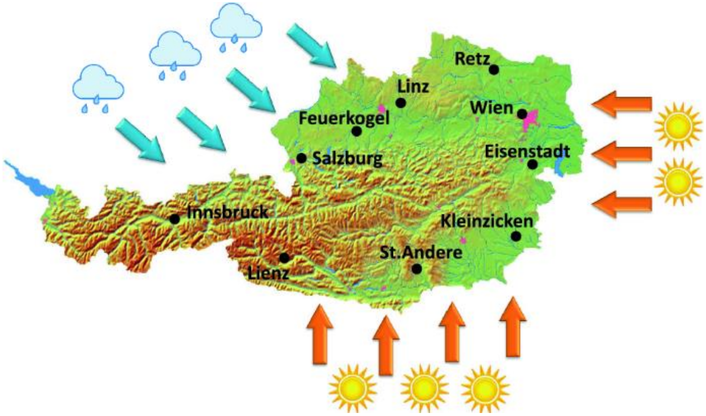
Team Data Mining

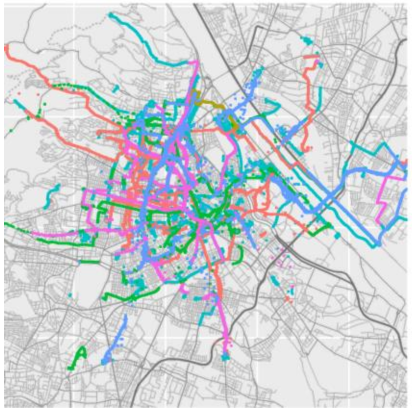
universität wien
Fakultät für Informatik

- Clustering
- Data mining methods for complex and heterogeneous data, e.g., time series, heterogeneous information networks
- Application-related methods: archeology, biomedicine, social sciences, meteorology, transport, particle physics









References

- D. Mautz, W. Ye, C. Plant, C. Böhm. Towards an Optimal Subspace for K-Means. KDD 2017
- D. Mautz, W. Ye, C. Plant, C. Böhm: Discovering Non-Redundant K-means Clusterings in Optimal Subspaces. KDD 2018
- L. Miklautz, D. Mautz, M. Altinigneli, C. Böhm, C. Plant: Deep Embedded Non-Redundant Clustering. AAAI 2020
- L. Miklautz, L. G. M. Bauer, D. Mautz, S. Tschiatschek: Details (Don't) Matter: Isolating Cluster Information in Deep Embedded Spaces. IJCAI 2021
- C. Leiber, D. Mautz, C. Plant, C. Böhm: Automatic Parameter Selection for Non-redundant Clustering. SDM 2022
- L. Miklautz, A. Shkabrii, C. Leiber, T. Bendeguz, B. Seidl, E. Weissensteiner, A. Rausch, C. Böhm, C. Plant. Non-redundant Image Clustering of Early Medieval Glass Beads. Under review (KDD 2023 Applications Track)

Step-by-step Transformation (for those interested ☺)

$$
\begin{aligned}
\mathcal{J} &= \left[\sum_{i=1}^{k}\sum_{x\in\mathcal{C}_i}\left\|P_{\mathrm{C}}^{\top}V^{\top}\mathbf{x}-P_{\mathrm{C}}^{\top}V^{\top}\boldsymbol{\mu}_i\right\|^2\right]+\sum_{\mathbf{x}\in\mathcal{D}}\left\|P_{\mathrm{N}}^{\top}V^{\top}\mathbf{x}-P_{\mathrm{N}}^{\top}V^{\top}\boldsymbol{\mu}_{\mathcal{D}}\right\|^2\\
&= \left[\sum_{i=1}^{k}\sum_{\mathbf{x}\in\mathcal{C}_i}\left(P_{\mathrm{C}}^{\top}V^{\top}\mathbf{x}-P_{\mathrm{C}}^{\top}V^{\top}\boldsymbol{\mu}_i\right)^{\top}\left(P_{\mathrm{C}}^{\top}V^{\top}\mathbf{x}-P_{\mathrm{C}}^{\top}V^{\top}\boldsymbol{\mu}_i\right)\right]\\
&\quad+\sum_{\mathbf{x}\in\mathcal{D}}\left(P_{\mathrm{N}}^{\top}V^{\top}\mathbf{x}-P_{\mathrm{N}}^{\top}V^{\top}\boldsymbol{\mu}_{\mathcal{D}}\right)^{\top}\left(P_{\mathrm{N}}^{\top}V^{\top}\mathbf{x}-P_{\mathrm{N}}^{\top}V^{\top}\boldsymbol{\mu}_{\mathcal{D}}\right)\\
&= \left[\sum_{i=1}^{k}\sum_{\mathbf{x}\in\mathcal{C}_i}(\mathbf{x}-\boldsymbol{\mu}_i)^{\top}VP_{\mathrm{C}}P_{\mathrm{C}}^{\top}V^{\top}(\mathbf{x}-\boldsymbol{\mu}_i)\right]+\sum_{\mathbf{x}\in\mathcal{D}}(\mathbf{x}-\boldsymbol{\mu}_{\mathcal{D}})^{\top}VP_{\mathrm{N}}P_{\mathrm{N}}^{\top}V^{\top}(\mathbf{x}-\boldsymbol{\mu}_{\mathcal{D}})
\end{aligned}
$$

Step-by-step Transformation cont. (for those interested ☺)

$$\mathcal{J} = \left[\sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} (\mathbf{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} V P_{\mathrm{C}} P_{\mathrm{C}}^{\mathsf{T}} V^{\mathsf{T}} (\mathbf{x} - \boldsymbol{\mu}_i) \right] + \sum_{\mathbf{x} \in \mathcal{D}} (\mathbf{x} - \boldsymbol{\mu}_{\mathcal{D}})^{\mathsf{T}} V P_{\mathrm{N}} P_{\mathrm{N}}^{\mathsf{T}} V^{\mathsf{T}} (\mathbf{x} - \boldsymbol{\mu}_{\mathcal{D}})$$

$$= \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \operatorname{Tr}\left((\mathbf{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} V P_{\mathrm{C}} P_{\mathrm{C}}^{\mathsf{T}} V^{\mathsf{T}} (\mathbf{x} - \boldsymbol{\mu}_i) \right) + \sum_{\mathbf{x} \in \mathcal{D}} \operatorname{Tr}\left((\mathbf{x} - \boldsymbol{\mu}_{\mathcal{D}})^{\mathsf{T}} V P_{\mathrm{N}} P_{\mathrm{N}}^{\mathsf{T}} V^{\mathsf{T}} (\mathbf{x} - \boldsymbol{\mu}_{\mathcal{D}}) \right)$$

$$= \operatorname{Tr}\left(P_{\mathrm{C}} P_{\mathrm{C}}^{\mathsf{T}} V^{\mathsf{T}} \left[\sum_{i=1}^{k} S_i \right] V \right) + \operatorname{Tr}\left(P_{\mathrm{N}} P_{\mathrm{N}}^{\mathsf{T}} V^{\mathsf{T}} S_{\mathcal{D}} V \right)$$

Step-by-step Transformation cont. (for those interested ☺)

$$\mathcal{J} = \operatorname{Tr}\left(P_{\mathrm{C}} P_{\mathrm{C}}^{\top} V^{\top} \left[\sum_{i=1}^{k} S_i\right] V\right) + \operatorname{Tr}\left(P_{\mathrm{N}} P_{\mathrm{N}}^{\top} V^{\top} S_{\mathcal{D}} V\right)$$

$$\Rightarrow \text{Use } \operatorname{Tr}(P_{\mathrm{N}}^{\top} A P_{\mathrm{N}}) = \operatorname{Tr}(A) - \operatorname{Tr}(P_{\mathrm{C}}^{\top} A P_{\mathrm{C}})$$

$$\mathcal{J} = \operatorname{Tr}\left(P_{\mathrm{C}}^{\top} V^{\top} \left[\sum_{i=1}^{k} S_i\right] V P_{\mathrm{C}}\right) - \operatorname{Tr}\left(P_{\mathrm{C}}^{\top} V^{\top} S_{\mathcal{D}} V P_{\mathrm{C}}\right) + \operatorname{Tr}\left(V^{\top} S_{\mathcal{D}} V\right)$$

$$= \operatorname{Tr}\left(P_{\mathrm{C}}^{\top} V^{\top} \underbrace{\left(\left[\sum_{i=1}^{k} S_i\right] - S_{\mathcal{D}}\right)}_{=:\Sigma} V P_{\mathrm{C}}\right) + \underbrace{\operatorname{Tr}\left(V^{\top} S_{\mathcal{D}} V\right)}_{\text{const. w.r.t. } V}$$